Work 20 (2003) 159–165
IOS Press

Speaking of Research

Pretest-posttest designs and measurement of change

Dimiter M. Dimitrov and Phillip D. Rumrill, Jr.
507 White Hall, College of Education, Kent State University, Kent, OH 44242-0001, USA
Tel.: +1 330 672 0582; Fax: +1 330 672 3737; E-mail: ddimitro@kent.edu

Abstract. The article examines issues involved in comparing groups and measuring change with pretest and posttest data. Different pretest-posttest designs are presented in a manner that can help rehabilitation professionals to better understand and determine effects resulting from selected interventions. The reliability of gain scores in pretest-posttest measurement is also discussed in the context of rehabilitation research and practice.

Keywords: Treatment effects, pretest-posttest designs, measurement of change

1. Introduction

Pretest-posttest designs are widely used in behavioral research, primarily for the purpose of comparing groups and/or measuring change resulting from experimental treatments. The focus of this article is on comparing groups with pretest and posttest data and related reliability issues. In rehabilitation research, change is commonly measured in such dependent variables as employment status, income, empowerment, assertiveness, self-advocacy skills, and adjustment to disability. The measurement of change provides a vehicle for assessing the impact of rehabilitation services, as well as the effects of specific counseling and allied health interventions.

2. Basic pretest-posttest experimental designs

This section addresses designs in which one or more experimental groups are exposed to a treatment or intervention and then compared to one or more control groups who did not receive the treatment. Brief notes on internal and external validity of such designs are first necessary. Internal validity is the degree to which the experimental treatment makes a difference in (or causes change in) the specific experimental settings. External validity is the degree to which the treatment effect can be generalized across populations, settings, treatment variables, and measurement instruments. As described in previous research (e.g. [11]), factors that threaten internal validity are: history, maturation, pretest effects, instruments, statistical regression toward the mean, differential selection of participants, mortality, and interactions of factors (e.g., selection and maturation). Threats to external validity include: interaction effects of selection biases and treatment, reactive interaction effect of pretesting, reactive effect of experimental procedures, and multiple-treatment interference. For a thorough discussion of threats to internal and external validity, readers may consult Bellini and Rumrill [1]. Notations used in this section are: Y_1 = pretest scores, T = experimental treatment, Y_2 = posttest scores, $D = Y_2 - Y_1$ (gain scores), and RD = randomized design (random selection and assignment of participants to groups and, then, random assignment of groups to treatments).

With the RDs discussed in this section, one can compare experimental and control groups on (a) posttest scores, while controlling for pretest differences or (b)

mean gain scores, that is, the difference between the posttest mean and the pretest mean. Appropriate statistical methods for such comparisons and related measurement issues are discussed later in this article.

Design 1: Randomized control-group pretest-posttest design

With this RD, all conditions are the same for both the experimental and control groups, with the exception that the experimental group is exposed to a treatment, T, whereas the control group is not. *Maturation* and *history* are major problems for internal validity in this design, whereas the *interaction of pretesting and treatment* is a major threat to external validity. Maturation occurs when biological and psychological characteristics of research participants change during the experiment, thus affecting their posttest scores. History occurs when participants experience an event (external to the experimental treatment) that affects their posttest scores. Interaction of pretesting and treatment comes into play when the pretest sensitizes participants so that they respond to the treatment differently than they would with no pretest. For example, participants in a job seeking skills training program take a pretest regarding job-seeking behaviors (e.g., how many applications they have completed in the past month, how many job interviews attended). Responding to questions about their job-seeking activities might prompt participants to initiate or increase those activities, irrespective of the intervention.

Design 2: Randomized Solomon four-group design

This RD involves two experimental groups, E_1 and E_2, and two control groups, C_1 and C_2. All four groups complete posttest measures, but only groups E_1 and C_1 complete pretest measures in order to allow for better control of pretesting effects. In general, the Solomon four-group RD enhances both internal and external validity. This design, unlike other pretest-posttest RDs, also allows the researcher to evaluate separately the magnitudes of effects due to treatment, maturation, history, and pretesting. Let D_1, D_2, D_3, and D_4 denote the gain scores for groups E_1, C_1, E_2, and C_2, respectively. These gain scores are affected by several factors (given in parentheses) as follows: D_1 (pretesting, treatment, maturation, history), D_2 (pretesting, maturation, history), D_3 (treatment, maturation, history), and D_4 (maturation, history). With this, the difference D_3–D_4 evaluates the effect of treatment alone, D_2–D_4 the effect of pretesting alone, and D_1–D_2–D_3 the effect of interaction of pretesting and treatment [11, pp. 68]. Despite the advantages of the Solomon four-group RD, Design 1 is still predominantly used in studies with pretest-posttest data. When the groups are relatively large, for example, one can randomly split the experimental group into two groups and the control group into two groups to use the Solomon four-group RD. However, sample size is almost always an issue in intervention studies in rehabilitation, which often leaves researchers opting for the simpler, more limited two-group design.

Design 3: Nonrandomized control group pretest-posttest design

This design is similar to Design 1, but the participants are not randomly assigned to groups. Design 3 has practical advantages over Design 1 and Design 2, because it deals with intact groups and thus does not disrupt the existing research setting. This reduces the reactive effects of the experimental procedure and, therefore, improves the external validity of the design. Indeed, conducting a legitimate experiment without the participants being aware of it is possible with intact groups, but not with random assignment of subjects to groups. Design 3, however, is more sensitive to internal validity problems due to interaction between such factors as selection and maturation, selection and history, and selection and pretesting. For example, a common quasi-experimental approach in rehabilitation research is to use time sampling methods whereby the first, say 25 participants receive an intervention and the next 25 or so form a control group. The problem with this approach is that, even if there are posttest differences between groups, those differences may be attributable to characteristic differences between groups rather than to the intervention. Random assignment to groups, on the other hand, equalizes groups on existing characteristics and, thereby, isolates the effects of the intervention.

3. Statistical methods for analysis of pretest-posttest data

The following statistical methods are traditionally used in comparing groups with pretest and posttest data: (1) Analysis of variance (ANOVA) on the gain scores, (2) Analysis of covariance (ANCOVA), (3) ANOVA on residual scores, and (4) Repeated measures ANOVA. In all these methods, the use of pretest scores helps to reduce error variance, thus producing more powerful tests than designs with no pretest data [22]. Generally speaking, the power of the test represents the probability of detecting differences between the groups being compared when such differences exist.

3.1. ANOVA on gain scores

The gain scores, $D = Y_2 - Y_1$, represent the dependent variable in ANOVA comparisons of two or more groups. The use of gain scores in measurement of change has been criticized because of the (generally false) assertion that the difference between scores is much less reliable than the scores themselves [5,14,15]. This assertion is true only if the pretest scores and the posttest scores have equal (or proportional) variances and equal reliability. When this is not the case, which may happen in many testing situations, the reliability of the gain scores is high [18,19,23]. The unreliability of the gain score does not preclude valid testing of the null hypothesis of zero mean gain score in a population of examinees. If the gain score is unreliable, however, it is not appropriate to correlate the gain score with other variables in a population of examinees [17]. An important practical implication is that, without ignoring the caution urged by previous authors, researchers should not always discard gain scores and should be aware of situations when gain scores are useful.

3.2. ANCOVA with pretest-posttest data

The purpose of using the pretest scores as a covariate in ANCOVA with a pretest-posttest design is to (a) reduce the error variance and (b) eliminate systematic bias. With randomized designs (e.g., Designs 1 and 2), the main purpose of ANCOVA is to reduce error variance, because the random assignment of subjects to groups guards against systematic bias. With nonrandomized designs (e.g., Design 3), the main purpose of ANCOVA is to adjust the posttest means for differences among groups on the pretest, because such differences are likely to occur with intact groups. It is important to note that when pretest scores are not reliable, the treatment effects can be seriously biased in nonrandomized designs. This is true if measurement error is present on any other covariate in case ANCOVA uses more than one (i.e., the pretest) covariate. Another problem with ANCOVA relates to differential growth of subjects in intact or self selected groups on the dependent variable [3]. Pretest differences (systematic bias) between groups can affect the interpretations of posttest differences.

Let us remind ourselves that assumptions such as randomization, linear relationship between pretest and posttest scores, and homogeneity of regression slopes underlie ANCOVA. In an attempt to avoid problems that could be created by a violation of these assumptions, some researchers use ANOVA on gain scores without knowing that the same assumptions are required for the analysis of gain scores. Previous research [4] has demonstrated that when the regression slope equals 1, ANCOVA and ANOVA on gain scores produce the same F ratio, with the gain score analysis being slightly more powerful due to the lost degrees of freedom with the analysis of covariance. When the regression slope does not equal 1, which is usually the case, ANCOVA will result in a more powerful test. Another advantage of ANCOVA over ANOVA on gain scores is that when some assumptions do not hold, ANCOVA allows for modifications leading to appropriate analysis, whereas the gain score ANOVA does not. For example, if there is no linear relationship between pretest and posttest scores, ANCOVA can be extended to include a quadratic or cubic component. Or, if the regression slopes are not equal, ANCOVA can lead into procedures such as the Johnson-Neyman technique that provide regions of significance [4].

3.3. ANOVA on residual scores

Residual scores represent the difference between observed posttest scores and their predicted values from a simple regression using the pretest scores as a predictor. An attractive characteristic of residual scores is that, unlike gain scores, they do not correlate with the observed pretest scores. Also, as Zimmerman and Williams [23] demonstrated, residual scores contain less error than gain scores when the variance of the pretest scores is larger than the variance of posttest scores. Compared to the ANCOVA model, however, the ANOVA on residual scores is less powerful and some authors recommend that it be avoided. Maxwell, Delaney, and Manheimer [16] warned researchers about a common misconception that ANOVA on residual scores is the same as ANCOVA. They demonstrated that: (a) when the residuals are obtained from the pooled within-group regression coefficients, ANOVA on residual scores results in an inflated α-level of significance and (b) when the regression coefficient for the total sample of all groups combined is used, ANOVA on residual scores yields an inappropriately conservative test [16].

3.4. Repeated measures ANOVA with pretest-posttest data

Repeated measures ANOVA is used with pretest-posttest data as a mixed (split-plot) factorial design with one between-subjects factor (the grouping vari-

Table 1
Pretest-posttest data for the comparison of three groups

Subject	Group	Pretest	Posttest	Gain
1	1	48	60	12
1	1	70	50	−20
1	1	35	41	6
4	1	41	62	21
5	1	43	32	−11
6	1	39	44	5
7	2	53	71	18
8	2	67	85	18
9	2	84	82	−2
10	2	56	55	−1
11	2	44	62	18
12	2	74	77	3
13	3	80	84	4
14	3	72	80	8
15	3	54	79	25
16	3	66	84	18
17	3	69	66	−3
18	3	67	65	−2

able) and one within-subjects (pretest-posttest) factor. Unfortunately, this is not a healthy practice because previous research [10,12] has demonstrated that the results provided by repeated measures ANOVA for pretest-posttest data can be misleading. Specifically, the F test for the treatment main effect (which is of primary interest) is very conservative because the pretest scores are not affected by the treatment. A very little known fact is also that the F statistic for the interaction between the treatment factor and the pretest-posttest factor is identical to the F statistic for the treatment main effect with a one-way ANOVA on gain scores [10]. Thus, when using repeated measures ANOVA with pretest-posttest data, the interaction F ratio, not the main effect F ratio, should be used for testing the treatment main effect. A better practice is to directly use one-way ANOVA on gain scores or, even better, use ANCOVA with the pretest scores as a covariate.

Table 1 contain pretest-posttest data for the comparison of three groups on the dependent variable Y. Table 2 shows the results from both the ANOVA on gain scores and the repeated measures ANOVA with one between subjects factor (Group) and one within subjects factor, Time (pretest-posttest). As one can see, the F value with the ANOVA on gain scores is identical to the F value for the interaction Group x Time with the repeated measures ANOVA design: $F(2, 15) = 0.56$, $p = 0.58$. In fact, using the F value for the between subjects factor, Group, with the repeated measures ANOVA would be a (common) mistake: $F(2, 15) = 11.34$, $p = 0.001$. This leads in this case to a false rejection of the null hypothesis about differences among the compared groups.

4. Measurement of change with pretest-posttest data

4.1. Classical approach

As noted earlier in this article, the classical approach of using gain scores in measuring change has been criticized for decades [5,14,15] because of the (not always true) assertion that gain scores have low reliability. Again, this assertion is true only when the pretest scores and the posttest scores are equally reliable and have equal variances. Therefore, although the reliability of each of the pretest scores and posttest scores should be a legitimate concern, the reliability of the difference between them should not be thought of as always being low and should not preclude using gain scores in change evaluations. Unfortunately, some researchers still labor under the inertia of traditional, yet inappropriate, generalizations. It should also be noted that there are other, and more serious, problems with the traditional measurement of change that deserve special attention.

First, although measurement of change in terms mean gain scores is appropriate in industrial and agricultural research, its methodological appropriateness and social benefit in behavioral fields is questionable. Bock [2, pp. 76] noted, “nor is it clear that, for example, a method yielding a lower mean score in an instructional experiment is uniformly inferior to its competitor, even when all of the conditions for valid experimentation are met. It is possible, even likely, that the method with the lower mean is actually the more beneficial for some minority of students”.

A second, more technical, problem with using raw-score differences in measuring change relates to the fact that such differences are generally misleading because they depend on the level of difficulty of the test items. This is because the raw scores do not adequately represent the actual ability that underlies the performance on a (pre- or post) test. In general, the relationship between raw scores and ability scores is not linear and, therefore, equal (raw) gain scores do not represent equal changes of ability. Fischer [7] demonstrated that, if a low ability person and a high ability person have made the same change on a particular ability scale (i.e., derived exactly the same benefits from the treatment), the raw-score differences will misrepresent this fact. Specifically, with a relatively easy test, the raw-score differences will (falsely) indicate higher change for the low ability person and, conversely, with a more difficult test, they will (falsely) indicate higher change for

Table 2
ANOVA on gain scores and repeated measures ANOVA on demonstration data

Model/Source of variation	*df*	*F*	*p*
ANOVA on gain scores			
Between subjects			
Group (G)	2	0.56	0.58
S within-group error	15		
Repeated measures ANOVA			
Between subjects			
Group (G)	2	11.34	0.00
S within-group error	15		
Within subjects			
Time (T)	1	5.04	0.04
T × G	2	0.56	0.58
T × G within-group error	15		

Note. The F value for G with the ANOVA on gain score is the same as the F value for T × G with repeated measures ANOVA.

the high ability person. Indeed, researchers should be aware of limitations and pitfalls with using raw-score differences and should rely on dependable theoretical models for measurement and evaluation of change.

4.2. Modern approaches for measurement of change

The brief discussion of modern approaches for measuring change in this section requires the definition of some concepts from classical test theory (CTT) and item response theory (IRT). In CTT, each observed score, X, is a sum of a *true score*, T, and *an error of measurement*, E (i.e., $X = T + E$). The true score is unobservable, because it represents the theoretical mean of all observed scores that an individual may have under an unlimited number of administrations of the same test under the same conditions. The statistical tests for measuring change in true scores from pretest to posttest have important advantages to the classical raw-score differences in terms of accuracy, flexibility, and control of error sources. Theoretical frameworks, designs, procedures, and software for such tests, based on structural equation modeling, have been developed and successfully used in the last three decades [13,21].

In IRT, the term ability connotes a latent trait that underlies performance on a test [9]. The ability score of an individual determines the probability for that individual to answer correctly any test item or perform a measured task. The units of the ability scale, called logits, typically range from -4 to 4. It is important to note that item difficulties and ability scores are located on the same (logit) scale. With the one-parameter IRT model (Rasch model), the probability of a correct answer on any item for a person depends on the difficulty parameter of the item and the ability score of the person. Fischer [7] extended the Rasch model to a Linear Logistic Model for Change (LLMC) for measuring both individual and group changes on the logit scale. One of the valuable features of the LLMC is that it separates the ability change into two parts: (1) *treatment effect*, the change part due to the experimental treatment, and (2) *trend effect*, the change part due to factors such as biological maturation, cognitive development, and other "natural trends" that have occurred during the pretest to posttest time period. Another important feature of the LLMC is that the two change components, treatment effect and trend effect, are represented on a ratio-type scale. Thus, the ratio of any two (treatment or trend) change effects indicates how many times one of them is greater (or smaller) than the other. Such information, not available with other methods for measuring change, can help researchers in conducting valid interpretations of change magnitudes and trends, and in making subsequent informed decisions. The LLMC has been applied in measuring change in various pretest-posttest situations [6,20]. A user-friendly computer software for the LLCM is also available [8].

5. Conclusion

Important summary points are as follows:

1. The experimental and control groups with Designs 1 and 2 discussed in the first section of this article are assumed to be equivalent on the pretest or other variables that may affect their posttest scores on the basis of random selection. Both designs control well for threats to internal and external validity. Design 2 (Solomon four-group design) is superior to Design 1 because, along with controlling for effects of history, maturation, and pretesting, it allows for evaluation of the magnitudes of such effects. With Design 3 (nonrandomized control-group design), the groups being compared cannot be assumed to be equivalent on the pretest. Therefore, the data analysis with this design should use ANCOVA or other appropriate statistical procedure. An advantage of Design 3 over Designs 1 and 2 is that it involves intact groups (i.e., keeps the participants in natural settings), thus allowing a higher degree of external validity.
2. The discussion of statistical methods for analysis of pretest-posttest data in this article focuses

on several important facts. First, contrary to the traditional misconception, the reliability of gain scores is high in many practical situations, particularly when the pre- and posttest scores do not have equal variance and equal reliability. Second, the unreliability of gain scores does not preclude valid testing of the null hypothesis related to the mean gain score in a population of examinees. It is not appropriate, however, to correlate unreliable gain scores with other variables. Third, ANCOVA should be the preferred method for analysis of pretest-posttest data. ANOVA on gain scores is also useful, whereas ANOVA on residual scores and repeated measures ANOVA with pretest-posttest data should be avoided. With randomized designs (Designs 1 and 2), the purpose of ANCOVA is to reduce error variance, whereas with nonrandomized designs (Design 3) ANCOVA is used to adjust the posttest means for pretest differences among intact groups. If the pretest scores are not reliable, the treatment effects can be seriously biased, particularly with nonrandomized designs. Another caution with ANCOVA relates to possible differential growth on the dependent variable in intact or self-selected groups.

3. The methodological appropriateness and social benefit of measuring change in terms of mean gain score is questionable; it is not clear, for example, that a method yielding a lower mean gain score in a rehabilitation experiment is uniformly inferior to the other method(s) involved in this experiment. Also, the results from using raw-score differences in measuring change are generally misleading because they depend on the level of difficulty of test items. Specifically, for subjects with equal actual (true score or ability) change, an easy test (a ceiling effect test) will falsely favor low ability subjects and, conversely, a difficult test (a floor effect test) will falsely favor high ability subjects. These problems with raw-score differences are eliminated by using (a) modern approaches such as structural equation modeling for measuring true score changes or (b) item response models (e.g., LLMC) for measuring changes in the ability underlying subjects' performance on a test. Researchers in the field of rehabilitation can also benefit from using recently developed computer software with modern theoretical frameworks and procedures for measuring change across two (pretest-pottest) or more time points.

References

[1] J. Bellini and P. Rumrill, *Research in rehabilitation counseling*, Springfield, IL: Charles C. Thomas.

[2] R.D. Bock, Basic issues in the measurement of change. in: *Advances in Psychological and Educational Measurement*, D.N.M. DeGruijter and L.J.Th. Van der Kamp, eds, John Wiley & Sons, NY, 1976, pp. 75–96.

[3] A.D. Bryk and H. I. Weisberg, Use of the nonequivalent control group design when subjects are growing, *Psychological Bulletin* **85** (1977), 950–962.

[4] I.S. Cahen and R.L. Linn, Regions of significant criterion difference in aptitude- treatment interaction research, *American Educational Research Journal* **8** (1971), 521–530.

[5] L.J. Cronbach and L. Furby, How should we measure change - or should we? *Psychological Bulletin* **74** (1970), 68–80.

[6] D.M. Dimitrov, S. McGee and B. Howard, Changes in students science ability produced by multimedia learning environments: Application of the Linear Logistic Model for Change, *School Science and Mathematics* **102**(1) (2002), 15–22.

[7] G.H. Fischer, Some probabilistic models for measuring change, in: *Advances in Psychological and Educational Measurement*, D.N.M. DeGruijter and L.J.Th. Van der Kamp, eds, John Wiley & Sons, NY, 1976, pp. 97–110.

[8] G.H. Fischer and E. Ponocny-Seliger, *Structural Rasch modeling*, Handbook of the usage of LPCM-WIN 1.0, Progamma, Netherlands, 1998.

[9] R.K. Hambleton, H. Swaminathan and H. J. Rogers, *Fundamentals of Item Response Theory*, Sage, Newbury Park, CA, 1991.

[10] S.W. Huck and R.A. McLean, Using a repeated measures ANOVA to analyze data from a pretest-posttest design: A potentially confusing task, *Psychological Bulletin* **82** (1975), 511–518.

[11] S. Isaac and W.B. Michael, *Handbook in research and evaluation* 2nd. ed., EdITS, San Diego, CA, 1981.

[12] E. Jennings, Models for pretest-posttest data: repeated measures ANOVA revisited, *Journal of Educational Statistics* **13** (1988), 273–280.

[13] K.G. Jöreskog and D.Sörbom, Statistical models and methods for test-retest situations, in: *Advances in Psychological and Educational Measurement*, D.N.M. DeGruijter and L.J.Th. Van der Kamp, eds, John Wiley & Sons, NY, 1976, pp. 135–157.

[14] L. Linn and J.A. Slindle, The determination of the significance of change between pre- and posttesting periods, *Review of Educational Research* **47** (1977), 121–150.

[15] F.M. Lord, The measurement of growth, *Educational and Psychological Measurement* **16** (1956), 421–437.

[16] S. Maxwell, H.D. Delaney and J. Manheimer, ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs, *Journal of Educational Statistics* **95** (1985), 136–147.

[17] G.J. Mellenbergh, A note on simple gain score precision, *Applied Psychological Measurement* **23** (1999), 87–89.

[18] J.E. Overall and J. A. Woodward, Unreliability of difference scores: A paradox for measurement of change, *Psychological Bulletin* **82** (1975), 85–86.

[19] D. Rogosa, D. Brandt and M. Zimowski, A growth curve approach to the measurement of change, *Psychological Bulletin* **92** (1982), 726–748.

[20] I. Rop, The application of a linear logistic model describing the effects of preschool education on cognitive growth, in: *Some*

mathematical models for social psychology, W.H. Kempf and B.H. Repp, eds, Huber, Bern, 1976.

[21] D. Sörbom, A statistical model for the measurement of change in true scores, in: *Advances in Psychological and Educational Measurement*, D.N.M. DeGruijter and L.J.Th. Van der Kamp, eds, John Wiley & Sons, NY, 1976, pp. 1159–1170.

[22] J. Stevens, *Applied multivariate statistics for the social sciences* 3rd ed., Lawrence Erlbaum, Mahwah, NJ, 1996.

[23] D.W. Zimmerman and R.H. Williams, Gain scores in research can be highly reliable, *Journal of Educational Measurement* **19** (1982), 149–154.